栽培大豆和野生大豆线粒体基因组密码子使用偏性的比较分析[1]

唐向民，杨守臻，陈怀珠，孙祖东，赖振光，曾维英，韦清源
（广西壮族自治区农业科学院 经济作物研究所，南宁 530007）

摘 要：为分析栽培和野生大豆线粒体基因组的密码子使用特征差异，该文以其线粒体基因组编码序列为研究对象，比较其密码子偏性形成的影响因素和演化过程。结果表明：(1)栽培大豆和野生大豆线粒体基因组编码区的 GC 含量分别为 44.56%和 44.58%，说明栽培大豆和野生大豆线粒体编码基因均富含 A/T 碱基。（2）栽培大豆和野生大豆线粒体基因组密码子第 1、第 2 位 GC 含量平均值与第 3 位 GC 含量的相关性均呈极显著水平，说明突变在其密码子偏性形成中的作用不可忽略；PR2-plot 分析显示，在同义密码子第 3 位碱基的使用频率上，嘌呤低于嘧啶；Nc-plot 分析中 Nc 比值位于-0.1~0.2 区间的基因数占总基因数的 95%以上；突变和选择等多重因素共同作用影响了大豆线粒体基因组编码序列密码子使用偏性的形成。(3)20、21 个密码子分别被确定为栽培大豆和野生大豆线粒体基因组编码序列的最优密码子，其中除丝氨酸 TCC 密码子外均以 A 或 T 结尾。综上结果显示，栽培大豆线粒体密码子偏性的形成受选择的影响高于野生大豆，这可能是栽培大豆由野生大豆经长期人工栽培驯化的结果。

关键词：栽培大豆，野生大豆，线粒体基因组，密码子偏性，最优密码子
中图分类号：S565.1　　**文献标识码：**A

Comparative analysis on codon usage bias in mitogenome of two species in genus *Glycine*

TANG Xiangmin, YANG Shouzhen, CHEN Huaizhu, SUN Zudong, LAI Zhenguang, ZENG Weiying, WEI Qingyuan
(*Cash Crops Research Institute, Guangxi Academy of Agricultural Sciences*, Nanning 530007, China)

Abstract: The characteristic of codon usage in mitogenome of wild and cultivated soybean was studied and compared to explore the main factors affecting codon usage bias and codon evolution.The results were as follows: (1) The GC content in the coding region of mitochondrial genome of cultivated soybean and wild soybean was 44.56% and 44.58% respectively, which indicated that the coding sequences of soybean mitogenome was rich in A/T base. (2) PR2-plot analysis showed that the third base of codon usage was biased, where pyrimidine was used more frequently than purine; Nc-plot analysis revealed that the number of genes with Nc ratio in the range of - 0.1-0.2 accounts for more than 95% of the total number of genes. The above-mentioned results showed that the multiple factors such as mutation and selection affected the formation of codon usage bias in soybean mitogenome. GC1&2 value was significantly correlated with GC3, suggesting that codon bias in the mitogenome of cultivated and wild soybeans was mainly affected by mutation. (3) In addition, 20 and 21 codons were identified to be the optimal codons for the mitogenome of cultivated and wild soybeans, with most of the preferred synonymous codons

---

[1]**收稿日期：2019-11-20**
基金项目：广西自然科学基金（2016GXNSFBA380127, 2014GXNSFAA118130）[Supported by Guangxi Natural Science Foundation （2016GXNSFBA380127, 2014GXNSFAA118130）]
作者简介：唐向民(1978 -)，女，广西玉林人，博士，助理研究员，主要从事大豆遗传育种研究，（E-mail）txm906@126.com。
通信作者：杨守臻，副研究员，主要从事大豆遗传育种研究，（E-mail）1336781130@189.cn。

ending with A or T base. The formation of mitochondrial codon bias of cultivated soybean was more affected by selection than that of wild soybean, which may be the result of long-term artificial cultivation and domestication of cultivated soybean.

Key words: *Glycine max*, *Glycine soja*, mitogenome, codon usage bias, optimal codon

同义密码子使用偏好性是指编码序列中同义密码子被生物体偏爱使用的不同频率（Hanson et al., 2018; Qiu et al., 2011）。由于生物界通用密码子的简并性，每个氨基酸至少对应 1 种密码子，最多有 6 种对应的密码子。同义的三联体密码子通常不是被随机使用的，这种偏好性确保了最优密码子可以与数量最多 tRNA 基因的反密码子配对（Sun et al., 2009）。同时避免了氨基酸的错掺，减少了翻译加工差错（Wei et al., 2014）。研究同义密码子使用的偏好性以及导致其形成的因素，可以帮助更好地了解生物基因组的特征、分子进化以及生态适应性等。之前关于陆地植物全基因组和叶绿体基因组密码子偏性的相关研究较多（Mazumdar et al., 2017），但植物线粒体基因组的密码子使用模式及其相关作用力尚未得到很好的研究。

作为一种细胞核外的遗传系统，线粒体基因组具有具有重排进化快、叶绿体 DNA 插入等结构特点。目前，对线粒体基因组的研究主要体现在对基因的结构和功能、基因表达的时空调控、核质互作、分子进化规律、物种的起源进化以及线粒体的起源等生物学领域的研究（夏玉玲等，2008；Kong et al., 2013）。 mtDNA 已被广泛应用于物种鉴定、种内母系演化、种间系统发育、种群系统进化及遗传多样性分析等方面的研究（Galtier et al., 2009; Wei et al., 2014; 农全东等, 2019; 李平等, 2019）。高等植物的线粒体基因组具有 RNA 编辑和密码子偏好性等特殊的表达方式（李玉秋等，2011）。了解植物线粒体基因组密码子的使用偏好，可以更好地了解其线粒体基因组进化过程。

大豆是豆科大豆属一年生草本植物。一年生野生大豆（*Glycine soja*）为栽培大豆（*Glycine. max*）的祖先种，是开展栽培大豆遗传育种工作的宝贵种质。栽培大豆（*G. max*）是从野生大豆(*G. soja*)经过人工改良驯化和长期定向选择逐渐积累有益农艺性状演化而成的。虽然栽培大豆和野生大豆的线粒体全基因组已完成测序（Sajjad et al., 2018, Chang et al., 2013），但关于两者的 mtDNA 编码基因的密码子使用模式比较的研究尚未见报道。比较研究野生大豆和栽培大豆 mtDNA 编码基因的密码子偏好性，能够更好地揭示两个同属物种的线粒体基因表达系统差异和进化差异。本研究拟以大豆属的野生大豆和栽培大豆线粒体基因组为研究对象，对其密码子使用特征进行系统分析，揭示影响其密码子偏性形成的主要因素，并确定最优密码子。通过比较两个同属物种对线粒体编码序列密码子使用的差异，以期为大豆在人工驯化进程中线粒体表达系统遗传变异的深入研究奠定理论基础。

1 研究方法

1.1 数据来源

在 NCBI 数据库检索一年生野生大豆和栽培大豆的线粒体基因组序列，其 GenBank 登录号分别为NC_039768.1 和NC_020455.1。分别下载2 个物种的线粒体基因组所有编码序列，剔除重复的编码序列以及小于 300 bp 的编码序列，并经开放阅读框研判，符合条件的余下编码序列用于本研究的偏性分析。

1.2 方法

1.2.1 中性绘图分析

分别统计野生大豆和栽培大豆线粒体基因组各CDS在密码子第1、第2和3位的GC含量，以GC3（第3位GC含量）为横坐标，以GC2（第2位GC含量）和GC1（第1位GC含量）的平均值为纵坐标（计为GC12）绘制二维散点图，以分析三联体密码子3 个位置碱基组成的相关性（Sueoka, 1988）。若GC12与GC3的相关性不显著，则表明密码子第1、第2位与第3位碱基使用存在差异，选择压力对密码子偏性影响较大；若GC12与GC3显著相关，表明密码子第1、第2位与第3位碱基使用无差异，密码子偏性的形成受突变的影响大。

1.2.2 相对同义密码子使用度分析

应用CodonW软件计算获得各编码基因的有效密码子数（Nc值）、密码子偏爱指数（Codon Bias Index, CBI）、最优密码子使用频率（Frequency of Optimal Codons，Fop）。利用CAIcal在线服务器对大豆线粒体基因组编码序列的相对同义密码子使用度（relative synonymous codon usage，RSCU）进行分析（Puigbo et al., 2008）。若RSCU=1，表明该密码子的使用无

偏好性；若RSCU>1，表明该密码子的使用频率大于同义密码子使用的平均频率；若RSCU<1，则表明低于平均频率。

1.2.3 Nc-plot 绘图分析

以GC3s为横坐标，Nc值为纵坐标，作散点图，以探讨碱基组成对密码子偏好性的影响；以仅由碱基组成决定密码子偏好性时的理论值作标准曲线，标准曲线反映了在突变压力下的Nc和GC3s的函数关系（Wright et al., 1990）。理论Nc值计算公式为$Nc=2+GC3s+29/[GC3s^2+(1-GC3s)^2]$。

1.2.4 奇偶偏好分析

PR2（parity rule 2，PR2）分析，为避免由密码子第3位碱基的A/T 或T/A 和G/C 或C/G 的突变不均衡，仅选择4种密码子编码的氨基酸，即丙氨酸、亮氨酸、脯氨酸、丝氨酸、苏氨酸、缬氨酸、精氨酸和甘氨酸，进行密码子第3位上4种碱基组成的分析。以每个基因A3/(A3+T3)值作纵坐标，G3/(G3+C3)值作横坐标，以平面散点图展示各基因碱基组成。中心点代表C=G且A=T，由中心点向坐标点发出的矢量则表示偏倚程度和方向（Sueoka, 1999）。

1.2.5 最优密码子分析

以 CDS 的 Nc 值为偏性标准，将大豆线粒体基因组编码基因中 Nc 值居于最低和最高两极的 10%基因分别构成高、低表达组，统计两组的 RSCU 值。当两组间△RSCU（高表达组 RSCU 值－低表达组 RSCU 值）大于 0.08 密码子定义为高表达密码子。将整体 RSCU >1 的密码子确定为高频率密码子。同时满足上述 2 种条件的密码子定义为最优密码子（Wang et al., 2018）。

2 结果与分析

2.1 密码子使用特征

栽培大豆和野生大豆在线粒体基因组编码区的碱基组成上基本一致，其总体GC含量分别为44.56%和44.58%。而它们的整个线粒体基因组GC含量均为45.03%。栽培大豆和野生大豆的线粒体编码基因在密码子3个不同位置的碱基GC含量也有所差异，密码子第一位碱基的GC 含量最高，分别为 48.27%、48.31%；第二位分别为 43.76%、43.52%；第三位分别为 41.64%、41.92%，呈现出GC1>GC2> GC3的趋势（表1）。也表明两个物种线粒体编码基因富含 A、T 碱基。

表 1 野生大豆和栽培大豆线粒体基因组密码子不同位置 GC 含量（单位：%）

Table 1 GC content of different positions of codon in *Glycine* mitogenome（Unit：%）

物种 Species	GCall	GCcds	GC1	GC2	GC3
栽培大豆 *G. max*	45.03	44.56	48.27	43.76	41.64
野生大豆 *G. soja*	45.03	44.58	48.31	43.52	41.92

栽培大豆和野生大豆的 Nc 与 GC3 分别呈显著相关和极显著相关，与 GC1 和 GC2 的相关性均未达显著水平，说明其线粒体编码基因密码子第 3 位上的碱基组成对密码子偏性影响较大。线粒体基因组编码序列的 GC1、GC2 和 GC3，在野生大豆中呈两两显著相关，表明其密码子第 1、第 2 和 3 位的碱基组成相似；而在栽培大豆中 GC1 与 GC2 及 GC3 的相关性均未达显著水平，仅 GC2 和 GC3 显著相关，表明其密码子第 1 位与第 2、第 3 位的碱基组成不同，第 2 和 3 位的碱基组成相似。在野生大豆中，Nc 还与线粒体基因组编码基因的 GC 含量（GCcds）及最优密码子使用频率（Fop）呈极显著相关，与 CBI 呈显著相关；而在栽培大豆中，Nc 与 GCcds 及 Fop 均不显著（表 2）。表明两种大豆属植物线粒体基因组密码子偏性的影响因素有所差异。

表 2 大豆线粒体基因密码子偏性相关参数的相关性

Table 2 The correlation of main parameters of gene codons in *Glycine* mtDNA

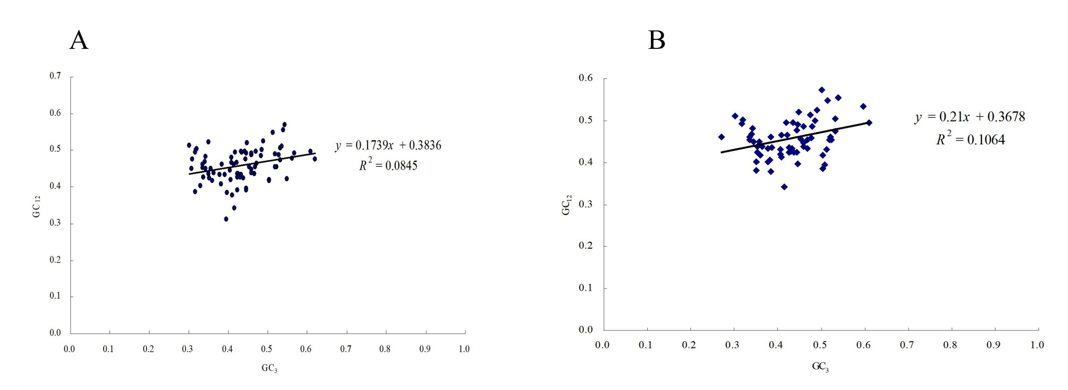
野生大豆	栽培大豆 *G. max*										
G. soja	GC1	GC2	GC3	Nc	GC3s	L_aa	GCcds	Gravy	CAI	CBI	Fop
GC1	-	0.125	0.187	0.136	0.204	0.176	0.571**	-0.239*	0.094	0.034	0.029
GC2	0.226*	-	0.247*	-0.126	0.272*	-0.063	0.680**	-0.105	0.170	0.408**	0.449**
GC3	0.292*	0.226*	-	0.264*	0.989**	-0.301**	0.766**	-0.222*	0.078	0.564**	0.566**
Nc	0.124	0.104	0.398**	-	0.267*	0.056	0.144	-0.214	-0.137	-0.009	0.008
GC3s	0.298**	0.234*	0.992**	0.410**	-	-0.303**	0.780**	-0.264*	0.077	0.584**	0.588**
L_aa	0.181	-0.060	-0.225*	0.042	-0.228*	-	-0.143	0.190	-0.116	-0.308**	-0.295**
GCcds	0.684**	0.683**	0.748**	0.308**	0.750**	-0.068	-	-0.276*	0.166	0.536**	0.555**
Gravy	-0.152	-0.210	-0.322**	-0.124	-0.348**	0.220*	-0.330**	-	0.311**	0.017	-0.095
CAI	0.182	-0.104	-0.007	0.104	-0.010	-0.028	0.025	0.232*	-	0.564**	0.564**
CBI	0.193	0.279*	0.455**	0.273*	0.478**	-0.165	0.449**	-0.104	0.435**	-	0.976**
Fop	0.169	0.327**	0.474**	0.290**	0.499**	-0.171	0.471**	-0.245*	0.458**	0.966**	-

注：**. 在 0.01 水平上显著相关；*. 在 0.05 水平上显著相关。

Note：One asterisk (*) and two asterisks (**) indicate correlations at a level of 0.05 and 0.01, respectively.

2.2 中性绘图分析

栽培大豆和野生大豆线粒体基因组各基因的中性绘图分析（图1），各基因坐标点未沿对角线趋势分布。栽培大豆线粒体CDS的GC12分布在0.3125~0.5680，GC3分布在0.3016~0.6190；一年生野生大豆线粒体CDS的GC12含量分布在0.3423~0.6045，GC3分布在0.2700~0.6095。两个物种GC12和GC3的相关系数分别为0.326和0.329，双侧检验表明其相关性达极显著水平；回归曲线斜率分别为0.1739和0.21。说明密码子第 1 、第 2 位与第 3 位碱基的碱基组成和变异模式相似，突变对基因密码子偏好性的形成有重要影响，其他因素（例如自然选择）可能对密码子偏好性形成的影响较小。栽培大豆线粒体密码子偏性的形成受选择的影响高于野生大豆，这可能是大豆长期人工栽培驯化的结果。



A. 栽培大豆；B. 野生大豆。GC12 表示密码子第1、第2 位碱基GC 平均含量；GC3 表示第3 位碱基GC 含量。

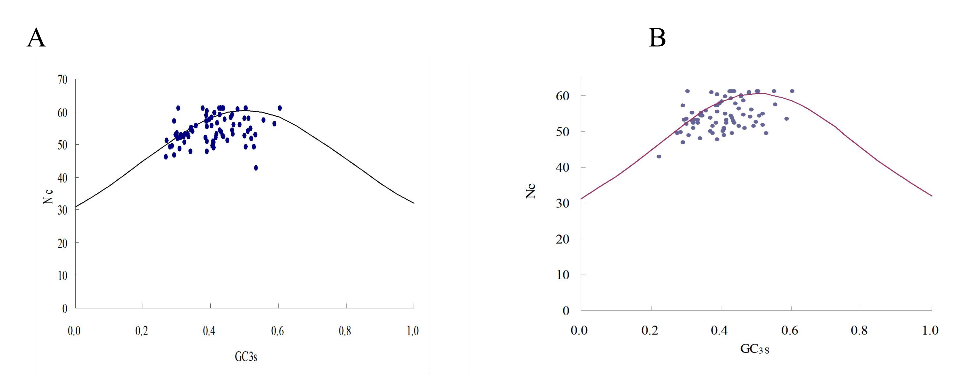
A. *G. max*; B. *G. soja*. GC_{12} means average GC content of the first and the second position of these codons, GC_3 means the GC content of the third position of codons.

图1 碱基组成对密码子偏性影响的中性绘图分析

Fig.1 Neutrality plot analysis of the effect of base composition on codon bias

2.3 Nc与GC3s 的关联分析

以GC3s为横坐标，各基因Nc值为纵坐标绘制Nc-plot曲线，以进一步判断突变或选择等因素对大豆线粒体基因组编码序列密码子偏好性形成的影响。根据（Nc预期－Nc实际）/Nc预期的公式计算Nc比值（表3）。栽培大豆和野生大豆Nc比值分布在－0.1~ 0.1 区间的线粒体基因占70%以上，这些基因对应的散点集中分布在标准曲线附近，其Nc实际值和预期值差异较小，表明这些线粒体基因的密码子偏好性的形成主要受突变因素的影响；而其余基因散点偏离于标准曲线下方（图2），其Nc实际值和预期值之间相差较大，表明这些基因的密码子偏好形成更多地受选择的影响。



A. 栽培大豆；B. 野生大豆。Nc表示有效密码子数; GC3s表示同义密码子第3位核甘酸的平均GC含量。
A. *G. max*; B. *G. soja*. Nc indicated effective number of condon; GC3s indicate the GC content of the third position of synonym codons.

图2 线粒体基因的Nc-plot分析

Fig.2 Nc-plot analysis on mitochondrial gene in *Glycine*

表3 Nc比值频率分布
Table 3 Distribution of Nc ration

组限 Class boundary	组中值 Class mid value	野生大豆 *G. soja* 组数 Frequency number	野生大豆 *G. soja* 频率 Frequency	栽培大豆 *G. max* 组数 Frequency number	栽培大豆 *G. max* 频率 Frequency
-0.20~0.10	-0.15	1	0.012	1	0.013
-0.10~0	-0.05	19	0.235	16	0.2
0~0.1	0.05	41	0.506	42	0.525
0.1~0.2	0.15	20	0.247	20	0.25
0.2~0.3	0.25	0	0	1	0.013
总数 Total		81	1	80	1

2.4 PR2-plot分析

采用PR2-plot方法分析栽培大豆和野生大豆线粒体基因组CDS的4种同义密码子第3位碱基组成（图3）。结果表明，散点并不是均匀地分布于PR2平面图中的4个区域内，50%以上散点位于平面图的左下区域，表明在同义密码子第3位碱基的使用频率上，A低于T，G低于C。若密码子使用偏好性仅由突变影响，则同义密码子第3位4种碱基的使用频率理论值相等，即A=T，C=G。A 和 T 以及 G 和 C 的使用频率不均衡性，表明两种大豆线粒体基因组密码子的使用偏好性由突变和选择等因素共同影响。

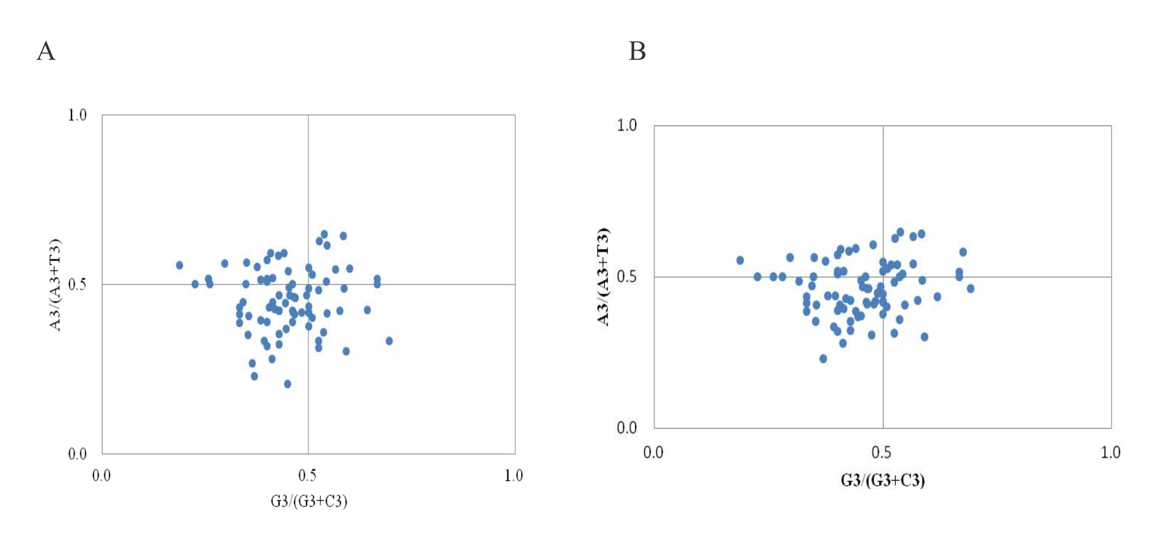


图3 栽培大豆（A）和野生大豆（B）线粒体基因的PR2-plot分析
Fig.3 PR2-plot analysis on mitochondrial gene in *G. max*（A）and *G. soja*（B）

2.5 最优密码子分析

对栽培大豆和一年生野生大豆线粒体基因组各编码序列的 Nc 值进行排序，分别从两极各选出 8 个基因构建高、低表达组。按组计算各基因同义密码子的 RSCU 值，并根据对高表达密码子和高频率密码子的定义，最终分别确定了栽培大豆和野生大豆 TTT、TTA、CTT、ATT、GTA、TCC、CCT、ACT、ACA、GCT、CAA、AAT、AAA、GAT、GAA、CGT、AGA、GGT 等 18 个共有的最优密码子。此外，确定了 CAT 和 CGA 为栽培大豆的最优密码子；GTT、TGT 和 GGA 为野生大豆的最优密码子。以上最优密码子中以 A 结尾有 9 个，以 T 结尾的有 13 个，以 C 结尾的有 1 个，无以 G 结尾的最优密码子。暗示了大豆线粒体基因组偏好于使用 A 或 T 结尾的密码子（表 4）。

表 4 大豆线粒体编码基因的相对同义密码子使用度

Table 4 The relative synonymous codon usage of mitochondrial genes in *Glycine*

氨基酸 Amino acid	密码子 Codon	栽培大豆 *G. max*			野生大豆 *G. soja*		
		基因组 Genome	高表达组 High experssion	低表达组 Low experssion	基因组 Genome	高表达组 High experssion	低表达组 Low experssion
Phe/F	TTT*	1.086	1.103	1.000	1.128	1.000	0.889
	TTC	0.914	0.897	1.000	0.872	1.000	1.111
Leu/L	TTA*	1.248	1.552	1.100	1.230	1.367	0.804
	TTG	0.996	1.007	1.050	1.050	1.063	1.237
	CTT*	1.362	1.385	1.300	1.398	1.557	1.361
	CTC	0.804	0.671	1.150	0.792	0.570	1.237
	CTA	0.936	0.671	0.800	0.888	0.835	0.804
	CTG	0.654	0.713	0.600	0.642	0.608	0.557
Ile/I	ATT*	1.254	1.514	1.017	1.221	1.387	1.109
	ATC	0.885	0.943	1.220	0.924	0.903	1.043
	ATA	0.861	0.543	0.763	0.855	0.710	0.848
Val/V	GTT*	1.208	1.278	1.254	1.156	1.151	1.018
	GTC	0.852	0.722	0.836	0.836	0.822	1.309
	GTA*	1.024	1.333	1.075	1.000	1.370	1.018
	GTG	0.920	0.667	0.836	1.008	0.658	0.655
Ser/S	TCT	1.368	1.477	1.455	1.338	1.422	1.620
	TCC*	1.140	1.385	0.864	1.140	1.378	1.051
	TCA	1.134	1.200	1.227	1.104	1.067	1.226
	TCG	0.774	0.554	0.864	0.840	0.622	0.745
	AGT	0.900	0.738	0.727	0.900	0.978	0.657
	AGC	0.684	0.646	0.864	0.678	0.533	0.701
Pro/P	CCT*	1.452	1.928	1.014	1.448	2.089	1.0625
	CCC	0.908	0.530	0.732	0.888	0.578	1.000
	CCA	1.012	1.012	1.352	1.024	0.889	1.125
	CCG	0.628	0.530	0.901	0.640	0.444	0.8125
Thr/T	ACT*	1.348	1.556	1.120	1.300	1.449	1.302
	ACC	1.072	0.500	1.360	1.048	0.464	1.302
	ACA*	1.060	1.278	1.040	1.096	1.449	1.116
	ACG	0.520	0.667	0.480	0.556	0.638	0.279
Ala/A	GCT*	1.504	2.025	1.500	1.456	2.085	1.222
	GCC	0.996	0.642	0.714	1.040	0.563	0.889
	GCA	0.928	0.790	0.929	0.936	0.958	0.778
	GCG	0.572	0.543	0.857	0.568	0.394	1.111
Tyr/Y	TAT	1.440	1.500	1.600	1.426	1.405	1.333
	TAC	0.560	0.500	0.400	0.574	0.595	0.667
His/H	CAT*	1.348	1.257	1.100	1.324	1.257	1.364
	CAC	0.652	0.743	0.900	0.676	0.743	0.636
Gln/Q	CAA*	1.244	1.421	1.059	1.288	1.318	1.091
	CAG	0.756	0.579	0.941	0.712	0.682	0.909
Asn/N	AAT*	1.286	1.375	0.950	1.276	1.235	0.971
	AAC	0.714	0.625	1.050	0.724	0.765	1.029
Lys/K	AAA*	1.168	1.127	0.903	1.168	1.228	1.148
	AAG	0.832	0.873	1.097	0.832	0.772	0.852
Asp/D	GAT*	1.258	1.389	1.200	1.234	1.450	1.286

	GAC	0.742	0.611	0.800	0.766	0.550	0.714
Glu/E	GAA*	1.314	1.467	1.414	1.278	1.381	1.320
	GAG	0.686	0.533	0.586	0.722	0.619	0.680
Cys/C	TGT*	1.154	1.043	1.000	1.154	1.412	1.200
	TGC	0.846	0.957	1.000	0.846	0.588	0.800
Arg/R	CGT*	0.906	0.975	0.857	0.900	1.091	0.835
	CGC	0.672	0.525	0.980	0.684	0.682	0.911
	CGA*	1.158	1.275	1.163	1.164	1.159	1.139
	CGG	0.846	0.825	0.735	0.804	0.955	1.063
	AGA*	1.446	1.800	1.163	1.446	1.432	1.215
	AGG	0.972	0.600	1.102	1.002	0.682	0.835
Gly/G	GGT*	1.216	1.846	0.928	1.168	1.746	1.100
	GGC	0.624	0.308	0.812	0.616	0.282	0.700
	GGA*	1.316	1.231	1.217	1.324	1.296	1.200
	GGG	0.848	0.615	1.043	0.892	0.676	1.000

注： *表示△RSCU≥0.08 且 RSCU 值≥1。

Note：* indicates △RSCU≥0.08 and RSCU value ≥1.

3. 讨论与结论

根据内共生理论，线粒体由α-蛋白质细菌祖先进化（Roger et al., 2017）。作为半自主型的细胞器，线粒体是细胞进行有氧呼吸的主要场所，其有自身的遗传表达系统，但大多数定位于线粒体的功能蛋白（>95%）是由核基因编码的（Giegé et al., 2005）。由于其含有相对独立的细胞质遗传信息，故线粒体基因组亦在揭示物种间亲缘关系和物种进化等方面具重要价值。通过比较核基因和线粒体基因在密码子使用模式上的差异，可以预测真核生物核糖体在细胞内的定位。此外，研究线粒体密码子的使用特征可以更好地了解线粒体基因组的进化过程及主要影响因素。

同义密码子的使用偏好性是物种长期演化的结果，形成这种偏好性的主要动力是自然选择和突变（Duret, 2002）。突变和选择等多重因素共同影响了大豆线粒体基因密码子使用偏好性的形成。比较栽培大豆和野生大豆的 ENC-plot、中性绘图和 PR2-plot 等分析，结果暗示了它们具有相似的密码子使用模式，这是因为两个物种在进化和亲缘关系上较近（Grantham, et al., 1980）。但两个物种在密码子使用模式上也存在一定的差异。按同一标准判断它们有相同的最优密码子，也有各自不同的最优密码子。

本研究确认了栽培大豆的 20 个和野生大豆的 21 个最优密码子。其中的 18 个最优密码子为两个物种共有，且多以 A 或 T 碱基结尾。栽培大豆和野生大豆的线粒体编码基因在密码子第 3 位的 GC 含量分别为 41.64%、41.92%。这与许多高等植物的叶绿体基因组第 3 位密码子偏好使用 A 或 T 碱基相似（尚明照等, 2011；傅建敏等，2017；原晓龙等，2019；Zhou et al., 2008）。而与线粒体基因不同，大豆的核基因最优密码子被确定为 UCC 和 GCC（张乐等，2011），暗示了大豆的核基因和线粒体基因在密码子使用偏性上的差异。按照同样的最优密码子计算方法，经比较分析，普通野生稻和栽培大豆共有的线粒体基因组最优密码子为 15 个，普通野生稻和野生大豆共有的线粒体基因组最优密码子为 16 个（金刚等，2019）。推测这并非偶然，而是植物线粒体基因组表达系统在进化中保守性的体现。尚需对其他高等植物的线粒体密码子进行归纳研究后得出结论。而非最佳密码子可能更多地与其二级结构相关（Chaney, 2017）。按照现代的植物细胞质雄性不育（Cytoplasmic male sterility，CMS）发生理论，细胞质雄性不育的发生与线粒体基因组的重组变异有关（Tang et al., 2017）。普通野生稻线粒体基因组中细胞质雄性不育相关基因 *CW-orf307* 与其线粒体基因组的密码子偏好性有所不同，TCT、CCT、ACA、TTT 和 CAT 等在线粒体基因组中被认定的最优密码子在 *CW-orf307* 基因中的使用频率并不高，暗示了 *CW-orf307* 独特的起源（金刚等，2019）。但野生大豆中的相关情况尚需对其线粒体基因组中的细胞质雄性不育基因进行鉴定并分析后才能确定。本研究通过分析比较野生大豆和栽培大豆线粒体基因组编码基因的密码子使用特征，将有助于进一步加深对大豆属植物的线粒体基因表达系统进化的理解。

参考文献：

CHANEY JL, STEELE A, CARMICHAEL R, et al., 2017. Widespread position-specific conservation of synonymous rare codons within coding sequences[J]. PLoS Comput Biol, 13(5): e1005531.

CHANEY JL, CLARK PL, 2015. Roles for synonymous codon usage in protein biogenesis[J]. Annu Rev Biophys, 44(1): 143-166.

CHANG S, WANG Y, LU J, et al., 2013. The mitochondrial genome of soybean reveals complex genome structures and gene evolution at intercellular and phylogenetic levels[J]. PLoS One, 8(2):e56502.

FU JM, SUO YJ, LIU HM, et al., 2017. Analysis on codon usage in the chloroplast protein-coding genes of *Diospyros* spp.[J]. Nonwood For Res, 35(2): 38-44. [傅建敏，索玉静，刘慧敏，等，2017. 柿属植物叶绿体蛋白质编码基因密码子用法[J]. 经济林研究，35(2)：38-44.]

GALTIER N, NABHOLZ B, GLEMIN S, et al., 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal[J]. Mol Ecol, 18(22): 4541-4550.

GIEG P, SWEETLOVE LJ, COGNAT V, et al., 2005. Coordination of nuclear and mitochondrial genome expression during mitochondrial biogenesis in *Arabidopsis*[J]. Plant Cell, 17(5): 1497-1512.

GRANTHAM R, GAUTIER C, GOUY M, et al., 1980. Codon catalog usage and the genome hypothesis[J]. Nucleic Acids Res, 8(1): 49-62.

HANSON G, COLLER J, 2018. Codon optimality, bias and usage in translation and mRNA decay[J]. Nat Rev Mol Cell Biol, 19(1): 20-30.

JIN G, WANG LP, LONG LY, et al., 2019. Analysis of codon usage bias in the mitochondrial protein-coding genes of *Oryza rufipogon*[J]. Plant Sci J, 37(2): 188-197. [金刚，王丽萍，龙凌云，等. 2019. 普通野生稻线粒体蛋白质编码基因密码子使用偏好性的分析[J]. 植物科学学报，37(2)：188-197.]

KONG FN, SUN PP, CAO M, et al., 2013. Complete mitochondrial genome of *Pyropia yezoensis*: reasserting the revision of genus *Porphyra*[J]. Mitochondrial DNA, 25(5): 335-226.

LI P, LIU YT, ZHAO ZX, et al., 2019. Study on the Origin of CMS Gene and DNA Barcode for Oryza Based on Rice Mitochondrial *atp6* Gene[J]. J Yunnan Agric Univ(Nat Sci Ed), 34(4): 555-563. [李平，刘雅婷，赵自仙，等，2019. 基于水稻线粒体 *atp6* 基因的稻属 CMS 基因起源及 DNA 条形码的研究[J]. 云南农业大学学报(自然科学)，34(4)：555-563.]

LI YQ, ZHAO HK, TAN H, et al. 2011. Analysis and comparison on characteristic of mitochondrial genome of eight plants[J]. Biotechnol Bull, 10):156-162. [李玉秋，赵洪锟，谭化，等，2011. 八种植物线粒体基因组结构特征分析与比较[J]. 生物技术通报，10：156-162.]

LIU HM, WUYUN TN, DU HY, 2016. Analysis of characteristic of codon usage in Ref gene of *Eucommia ulmoides*[J]. J Cent S Univ For & Tech , 36(3):8-12. [刘慧敏，乌云塔娜，杜红岩，2016. 杜仲 *Ref* 基因密码子偏好性分析[J]. 中南林业科技大学学报，36(3)：8-12.]

MAZUMDAR P, BINTI OTHMAN R, MEBUS K, et al., 2017. Codon usage and codon pair patterns in non-grass monocot genomes[J]. Ann Bot, 120(6): 893-909.

NONG QD, ZHANG MY, JIAO ZL, et al., 2019. Cloning and potential application analysis of mitochondrial *Cytochrome b* gene in dragon fruit[J]. Mol Plant Breeding, 17(13):4194-4203. [农全东，张明永，焦正利，等，2019. 火龙果线粒体细胞色素 b 基因克隆及其应用潜力分析[J]. 分子植物育种，17(13)：4194-4203.]

PUIGBO P, BRAVO IG, GARCIA-VALLVE S, 2008. CAIcal: a combined set of tools to assess

codon usage adaptation[J]. Biol Direct, 3:38.

QIU S, ZENG K, SLOTTE T, et al., 2011. Reduced efficacy of natural selection on codon usage bias in selfing *Arabidopsis* and *Capsella* species[J]. Genome Biol Evol, 2011, 3: 868-880.

ROGER AJ, MUOZ-G MEZ SA, KAMIKAWA R, 2017. The origin and diversification of mitochondria[J]. Curr Biol, 27(21): R1177-R1192.

SAJJAD A, LATIF KA, AHMED AH, et al., 2018. The first complete mitochondrial genome of wild soybean (*Glycine soja*)[J]. Mitochondrial DNA Part B, 3(2):527-528.

SUEOKA N, 1999. Translation-coupled violation of Parity Rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position[J]. Gene(Amsterdam), 238(1): 0-58.

SUEOKA N, 1988. Directional mutation pressure and neutral molecular evolution[J]. Proc Natl Acad Sci USA, 85(8): 2653-2657.

SUN Z, WAN D, MURPHY RW, et al., 2009. Comparison of base composition and codon usage in insect mitochondrial genomes[J]. Genes Genom, 31(1):65-71.

TANG H, XIE Y, LIU YG, et al. Advances in understanding the molecular mechanisms of cytoplasmic male sterility and restoration in rice[J]. Plant Reprod, 2017, 30(4):179-184.

WANG B, LIU J, JIN L, et al., 2010. Complex mutation and weak selection together determined the codon usage bias in bryophyte mitochondrial genomes[J]. J Integr Plant Biol, 52(12): 1100-1108.

WANG L, XING H, YUAN Y, et al., 2018. Genome-wide analysis of codon usage bias in four sequenced cotton species[J]. PLoS One, 13(3): e0194372.

WEI L, HE J, JIA X, et al., 2014. Analysis of codon usage bias of mitochondrial genome in *Bombyx moriand* its relation to evolution[J]. BMC Evol Biol, 14(1): 262.

WEI SJ, LI Q, VAN ACHTERBERG K, et al., 2014. Two mitochondrial genomes from the families *Bethylidae* and *Mutillidae*: Independent rearrangement of protein-coding genes and higher-level phylogeny of the Hymenoptera[J]. Mol Phylogenet Evol, 77:1–10.

WRIGHT F, 1990. The 'effective number of codons' used in a gene[J]. Gene, 87(1): 23–29.

XIA YL, 2003. the study on mitochondrial genome of *B.mandarina* in China[D]. Chongqing: Southwest Agircultural University, 2-10. [夏玉玲，2003. 中国野桑蚕线粒体基因组研究[D]. 重庆：西南农业大学， 2-10.]

YUAN XL, LI YQ, ZHANG JF, et al., 2019. Analysis of codon usage bias in the chloroplast genome of *Dalbergia odorigera*[J]. Guihaia, 2019-11-05. [原晓龙，李云琴，张劲峰，等，2019. 降香黄檀叶绿体基因组密码子偏好性分析[J]. 广西植物，2019-11-05]

ZHANG L, JIN LG, LUO L, et al., 2011. Analysis of nuclear gene codon bias on soybean genome and transcriptome**[J]**. Acta Agr Sin, 37(6): 965-974. [张乐，金龙国，罗玲，等，2011. 大豆基因组和转录组的核基因密码子使用偏好性分析[J]. 作物学报，37(06)：965-974.]

ZHOU M, LI X, 2009. Analysis of synonymous codon usage patterns in different plant mitochondrial genomes[J]. Mol Biol Rep, 36(8): 2039-2046.

ZHOU M, LONG W, LI X, 2008. Patterns of synonymous codon usage bias in chloroplast genomes of seed plants[J]. Fores StudChina, 10(4): 235-242.